基于改进 GPT 模型的文本生成研究

王蛟 李慧
首都师范大学 教育学院，北京 100048

摘要：

[目的] 本研究旨在提出一种基于词和词性的联合文本生成模型，以提高生成文本的质量。

[方法] 该模型由两个预训练的文本生成模型组成，一个是基于词的模型，另一个是基于词性的模型。此外，本文还提出并使用了 BERT 模型对进行二分类任务，以判断文本生成效果。

[结果] 在三个数据集上的实验结果表明，与传统的 GPT 模型相比，GPT-WP 模型生成文本的质量有明显提升。

[局限] BERT 模型在二分类任务中参数较大，大规模数据训练下评价效果差，本文提出的模型在数据量较小的场景下表现较好，大规模数据表现差异缩小。

[结论] GPT-WP 模型在本文提出的评价方法下表明其能够有效地提高生成文本的质量，对于自然语言生成任务的改进和评估提供了参考。

关键词：联合生成模型　文本生成　BERT 模型　NLTK　评价指标

分类号：TP391

Content Generation Research Based on Improved GPT Model

WANG Jiao，LI Hui
(College of Education, Capital Normal University, Beijing 100048, China)

Abstract:

[Objective] This study aims to propose a joint text generation model based on words and lexicality to improve the quality of generated text.

[Methods] The model consists of two pre-trained text generation models, one is a word-based model and the other is a lexical-based model. In addition, the BERT model is proposed and used in this paper for performing a dichotomous classification task to judge the text generation effect.

[Results] Experimental results on three datasets show that the GPT-WP model generates text with significantly higher quality compared to the traditional GPT model.

[Limitations] The BERT model has larger parameters in the binary classification task and is poorly evaluated under large-scale data training. The model proposed in this paper performs better in scenarios with smaller amounts of data, and the difference in performance is reduced for large-scale data.

[Conclusions] The GPT-WP model is shown to be effective in improving the quality of generated text under the evaluation method proposed in this paper, which provides a reference for the improvement and evaluation of natural language generation tasks.

Keywords: Joint Generation Model; Text Generation; BERT Model; NLTK; Evaluation Indicators

1 引言

近年来，随着深度学习和自然语言处理技术的飞速发展，文本生成模型已经引起了广泛的关注和研究，例如文章生成[1]、诗词创作[2]、新闻自动编写[3]、智能对话系统[4]等。相比基于规则的文本生成技术[5]，基于深度学习的文本生成模型具备更多的技术优势,但也有更多的挑战，深度学习生成模型需要解决上下文信息长距离依赖的问题[6]、语义的连贯性[7]和多样性问题[8]等。目前，文本生成模型的研究主要集中在如何从复杂的上下文和文本数据中提取有效的语义信息并用于生成新的文本。

最新的文本生成方法多关注于序列学习、语义理解及上下文关系推理等。例如,Sutskever 等人利用深度神经网络[8](DNN)对文本生成中的序列学习进行建模,但是这类方法存在长距离信息传递的梯度消失和梯度爆炸问题。Vaswani 等人构建的 Transformer 模型[9]采用 Encoder-Decoder 结构，相较于前述模型，其在文本生成方面表现更加优异。

目前主流的文本生成模型广泛采用基于 Transformer 模型的 Encoder 和 Decoder 分别构建的 GPT 模型[10]和 BERT 模型[11]，其对文本进行分词处理，以词为单位对文本进行训练，并在生成任务中以 token 为单位输出单词或标点符号。

本文认为，目前的文本生成模型研究中，在提升模型规模的同时，可以通过对语法结构的学习与调控，实现对文本生成模型的优化，使其在较小的模型下，实现较好的文本生成结果。本文基于 GPT 模型开发了一种采用两个 GPT 模型进行联合预测的新模型 GPT-WP（Generative Pre-trained Transformer for Word and Part of speech)，着重对文本中词性的部分进行处理，采用一个参数量较小的模型对词性规律进行学习，另一个参数量较大的模型对词进行学习，并采用两个模型联合预测以改善生成结果，使生成文本具备更好的语法结构，更加贴近人类撰写文本的语法规律。

2 相关工作

2.1 LSTM

长短期记忆（LSTM）是循环神经网络（RNN）的一种，旨在解决梯度消失的问题。LSTM 能够有选择地更新或丢弃记忆中的信息，以使其具备不受梯度消失影响的能力。LSTM 和传统的 RNN 之间的主要区别是增加了一个记忆单元，其负责存储信息[6]。这个记忆单元由三个门控制：输入门、遗忘门和输出门。输入门决定记忆单元添加的信息，遗忘门决定从记忆单元中丢弃的信息，输出门决定从记忆单元中输出的信息。LSTM 的结构图，如图 1 所示。

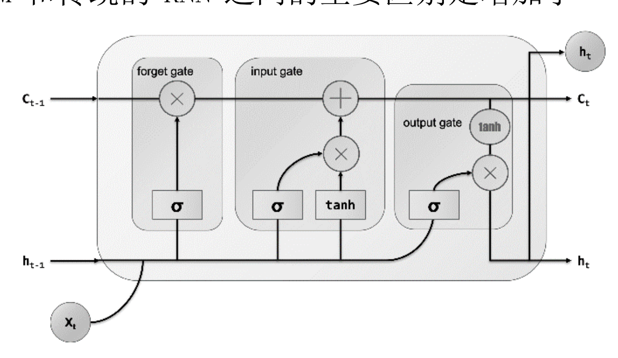


图 1 LSTM 结构图

输出C_{t-1}为上一神经元传递的记忆信息，h_{t-1}为上一神经元传递的状态信息，

X_t为传入的数据，输出为本神经元的记忆信息C_t和状态信息h_t并将其传递给下一神经元，其中的公式如下：

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma\left(W_{fx}x_t + W_{fh}h_{t-1} + b_f\right) \quad (2)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \quad (4)$$

$$h_t = o_t \odot \tanh(c_t) \quad (5)$$

LSTM 模型很难解释门和存储单元的行为，这种不可解释性使调试和改进模型的难度增大。此外，LSTM 的训练计算成本很高，其成本随数据集规模增大而增加。

LSTM 模型很难解释门和存储单元的行为，这种不可解释性使调试和改进模型的难度增大。此外，LSTM 的训练计算成本很高，其成本随数据集规模增大而增加。

2.1 Transformer

Transformer 最初是由 Vaswani 等人在 2017 年的开创性论文“Attention Is All You Need”中提出的[9]。Transformer 最初是为机器翻译任务设计的，但目前已广泛应用于各种自然语言处理任务，包括语言建模、问题回答和总结等。

Transformer 基于自注意力机制并行处理输入，颠覆了传统的递归神经网络[12]（RNN）和卷积神经网络[13]（CNN）中顺序处理的方式。自注意力机制通过计算所有输入标记对之间的权重，来衡量每个标记对网络输出的贡献，从而捕捉标记之间的关系，相比 RNN 和 CNN 能更有效地模拟长距离依赖关系。

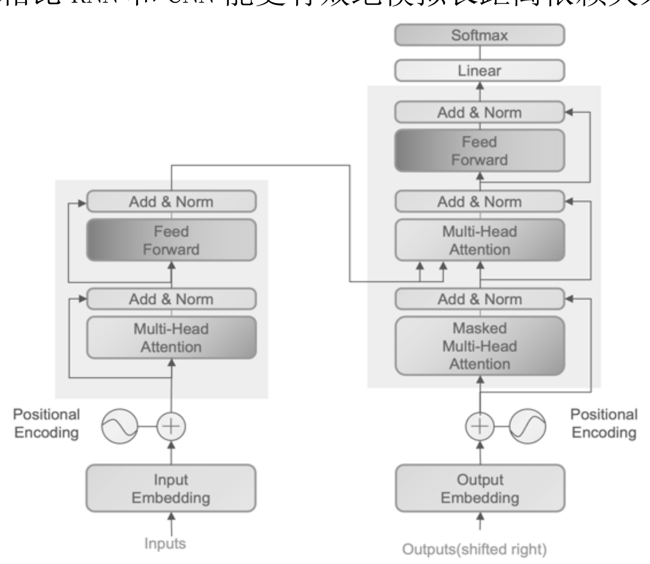


图 2 Transformer 结构图

Transformer 由一个编码器和一个解码器组成，每个编码器包含一个自注意力层和一个前馈层。编码器处理输入序列并产生一个隐藏状态的序列，然后将其作为输入给解码器。解码器使用编码器的隐藏状态和先前生成的标记作为输入，一次生成一个输出序列，如 Transformer 的结构图图 2 所示。

2.2 BERT

BERT（Bidirectional Encoder Representations from Transformers，来自 Transformer 的双向编码器表征）是由 Devlin 等人在 2018 年提出的基于 Transformer 架构的预训练语言模型[11]。其在大量无标注的文本数据上进行训练，使其能够学习高质量的自然语言表示。

BERT 结构图如图 3 所示，BERT 模型采用 Transformer 中的 Encoder 结构，其采用的自注意力机制对文本序列进行处理并预测，可实现下句预测（NSP）等任务。

BERT 的主要特征是双向性，即其在生成一个词的表示时能够考虑到该词的上下文，同时考虑左边和右边的语境，从而产生更准确的语言表征。

BERT 通常用于两个阶段的过程：预训练和微调。在预训练期间，BERT 在大量的未注释文本数据上进行训练，可以完成掩码语言模型（MLM）任务和下句预测（NSP）任务。MLM 任务使用随机屏蔽输入序列中一定比例的标记获得数据，并要求模型预测缺失的标记，而 NSP 任务包括预测原始文本中两个句子是否连续等任务。

在预训练之后，可以通过添加特定任务的输出层和在该任务的注释数据上训练模型，为情感分析或文本分类等下游任务进行微调。以这种方式对 BERT 进行微调已被证明可以在广泛的自然语言处理任务中取得较好结果。

BERT 的一个主要优势是它能够学习高质量的语言表征，并可以针对广泛的下游任务进行微调；另一个优势是它能够对一个词的上下文进行双向学习，从而提高其表征的准确性。BERT 模型目前已被广泛用于情感分析、命名实体识别和问题回答等场景。BERT 模型在预训练和微调阶段需要大量的计算资源，同时其模型存在解释困难等问题，也使 fine-tune 等调试工作难度增加。

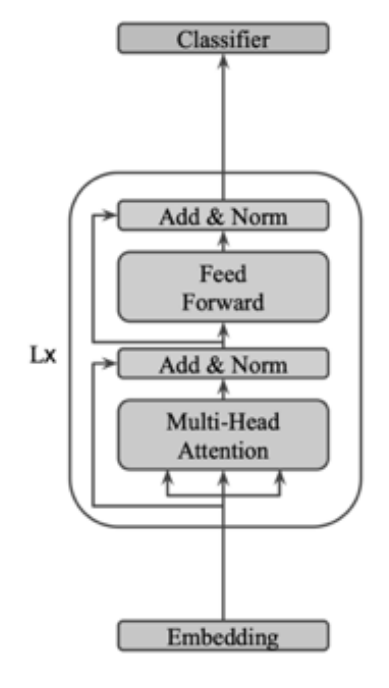


图 3 BERT 结构图
Fig.3 Structure of BERT

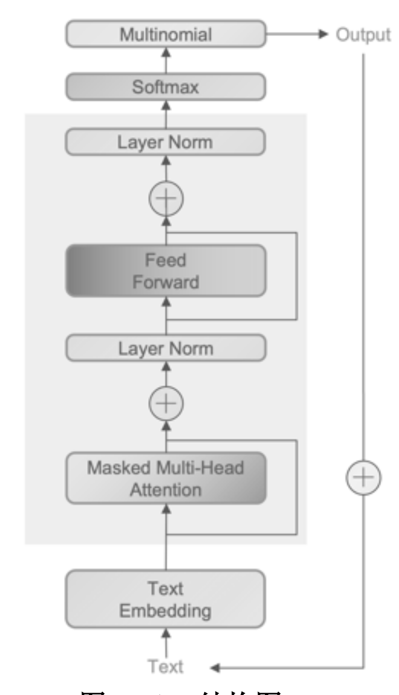


图 4 GPT 结构图
Fig.4 Structure of GPT

2.3 GPT

GPT 的基本结构由多层 Transformer 中的 Decoder 结构组成，每个 Decoder 包括一个注意力机制（Masked Multi-head Attention）和一个全连接前馈网络（Feed Forward Network），如图 3 所示。其中，注意力机制是指每个输入位置的注意力权重都由该位置和其他所有位置计算得到，这样可以同时考虑句子中所有位置的信息。注意力权重用于计算输入向量的加权和，经全连接网络处理后传入下一步骤。

每个 Decoder 的输出被以相同的方式，作为输入传入到下一个 Decoder 中。GPT 模型的最后一层是一个 Softmax 层，进行归一化，将最终隐藏状态映射到词汇表大小的向量中，在概率组中抽取结果作为下一层的输入（本文中模型采取 multinomial 的方法）。

2.4 nanoGPT

nanoGPT[14]是 GPT 模型的一个轻量级实现版本，主要用于资源受限的设备上进行文本生成任务。nanoGPT 采用了 GPT-2 的基本结构，但在其实现过程中，其采用在 Embedding 层后接入 Layer Normalization 层，并在 Feed Forward 网络后计算加权和采用 Softmax 方法进行归一，再通过 Multinomial 方法抽取出输出，如图 5 所示。

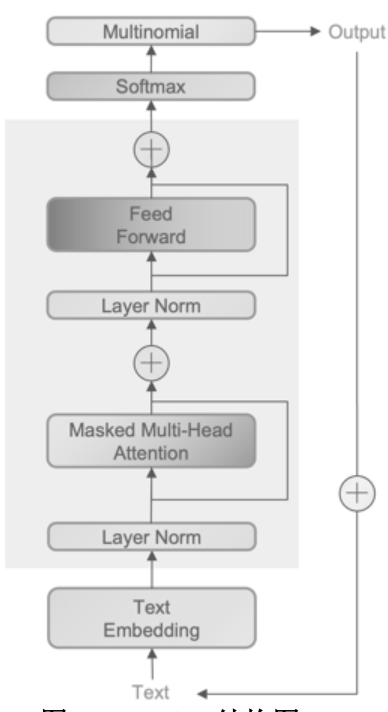


图 5 nanoGPT 结构图

3 本文方法

3.1 GPT-WP 模型

本研究提出了一个基于词和词性的联合生成文本模型 GPT-WP(Word POS)，由两个预训练的文本生成模型组成，一个是基于词的模型，另一个是基于词性的模型。在该模型中，输入文本和通过 NLTK 标记出的词性序列分别编码并传递给两个模型，如图 6 所示。

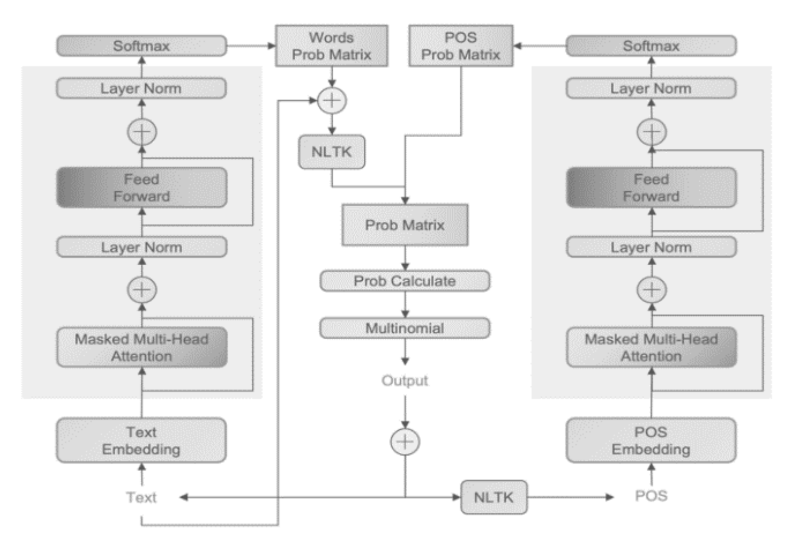


图 6 GPT-WP 结构图

为了使两模型可以协调生成模型，预防语法模型主导文本生成，本研究所提出的模型间采用残差相乘方法进行连接，公式如下：

$$P = P_{word} * (1 + P_{POS}) \tag{6}$$

其中，P为 GPT-WP 模型得出的最终概率，P_{word}为词模型输出的概率，P_{POS}为词性模型输出的概率。

在生成新的单个单词时，程序将其概率值与其对应的词性标记的概率值进行残差相乘，其中词性标记的概率值P_{POS}由词性模型生成，词的概率值P_{word}由词模型生成。然后程序从加权相乘后得到的概率值P中进行随机抽取，并以此生成最终的单个单词。通过将生成的单个单词添加到已生成的文本中，并将其作为新的输入文本传递给下一个循环迭代，逐步生成最终的文本，具体流程如图 6 和图 7 所示。

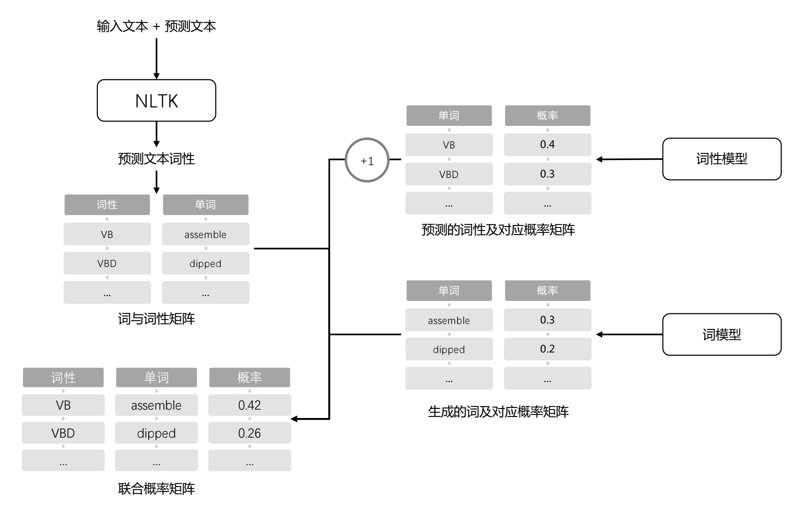


图 7 GPT-WP 模型中两模型联合计算流程图

3.2 评价方法

本文的评估目的在于对基于 GPT 和 GPT-WP 的文本生成数据进行评估，并比较它们在文本生成质量方面的性能。本研究使用 NLTK 工具对 CNN Dailymail 数据进行分词处理，获得人工数据和输入数据，并分别使用基于 GPT 的传统模型和

新模型对输入数据进行训练和预测，获得两个模型所对应的传统 GPT 模型数据和 GPT-WP 模型数据，具体结构如图 8 所示。

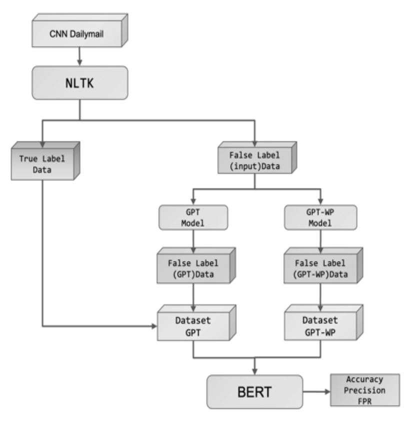


图 8 评价方法

1) 从 CNN Dailymail 数据集中随机选取部分样本，通过 NLTK 进行分句，保留每个样本的前 2 句，组成新的数据集。
2) 选取新数据集的一半作为人工数据(True Label Data)，将另一半数据集中第二句话的后 75%去除，作为文本生成模型的输入(False Label (input) Data)。
3) 将上述输入数据(False Label (input) Data)通过传统的 GPT 模型和 GPT-WP 模型进行预测，对预测结果进行截断，获取前两句作为输出，并得出对应的传统 GPT 模型生成数据(False Label(GPT) Data)和 GPT-WP 模型生成数据(False Label(GPT-WP) Data)。
4) 将上述传统 GPT 模型生成数据(False Label(GPT) Data)和 GPT-WP 模型生成数据(False Label(GPT-WP) Data)与人工数据(True Label Data)结合，形成传统 GPT 模型数据集(Dataset Original)和 GPT-WP 模型数据集(Dataset GPT-WP)
5) 分别将上述两数据集通过分割划分为训练集和测试集，通过 BERT 模型进行训练，得出对应的评价指标。

本研究采用所述评估方法对生成文本的质量进行评估，具体实施方式是使用 BERT 模型对原始数据和生成数据进行二元分类任务，以判断文本是否为人工编写的。评估过程涉及计算模型的各项性能指标，如准确率和精确率，从而评估其对生成文本质量评估的能力。以下是相关的公式：

$$Acc = \frac{Correct}{All} \tag{7}$$

其中，Acc为准确率，$Correct$为测试样本中预测正确的数量，All为测试样本总数。

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$FPR = \frac{FP}{TN + FP} \tag{9}$$

其中，P为精确率，TP为预测正确的人工数据个数，FP为预测错误的生成数据个数，TN为预测正确的生成数据个数。

准确率Acc衡量了 BERT 模型对于生成文本与原始数据正确分类的能力，精确度P指标反映了 BERT 模型将生成文本正确分类为人工撰写的能力，假正例率

FPR 通过计算生成数据中被判定为人工数据的比例，直接反映出模型所生成数据与人工数据的接近程度。

当 BERT-tiny 模型经训练后在测试集上的精确率P高时，表明 BERT-tiny 模型对人工文本寻找特征较多，在 BERT-tiny 所学习的维度上，能更加准确学习人工文本的特征， BERT-tiny 表现好，生成数据与人工数据共同特征较少，表现更差。

当 BERT-tiny 模型经训练后在测试集上的精确率P低时，表明 BERT-tiny 模型对人工的文本寻找特征较少，在 BERT-tiny 所学习的维度上，难以准确学习人工文本的特征， BERT-tiny 表现差，生成数据与人工数据更加接近。

当 BERT-tiny 模型经训练后在测试集上的假正例率FPR低时，表明 BERT-tiny 模型对生成的文本寻找特征较多，具备较明显的生成特征，其与人工数据的特征不符的更多，BERT 更易区分生成文本与人工文本，在 BERT 所学习的维度上，生成的文本与人工文本更接近，生成文本效果较差。

当 BERT-tiny 模型经训练后在测试集上的假正例率FPR高时，表明 BERT-tiny 模型在分辨生成文本与人工文本时产生更大困难，被标记为的正例较少，意味着模型将更多的负例错误地预测为正例，表明在 BERT 模型的视角下，,难以准确分辨，其具备更多的能被 BERT 模型寻找到的生成特征，生成的文本与人工数据更加接近，生成文本效果较好。

考虑准确率 *Acc* 的评估时，本研究结合精确率 *P* 和假正例率 *FPR* 进行分析。当精确率 *P* 较低、假正例率 *FPR* 较高且准确率 *Acc* 下降或变化不显著时，表示被正确标记为真实的正例较少。这意味着模型将更多的负例错误地预测为正例，从 BERT-tiny 模型的视角来看，生成文本与人工文本更加接近。反之，如果评估指标显示相反的结果,则表明生成文本与人工文本的差异较大,模型生成效果不佳。

综上，本研究可以得出，当 BERT 模型的分类效果相对良好时，表明其生成的数据与真实数据的差异相对较小，模型效果相对较好。相反，当 BERT 模型的分类效果相对差时，表明其生成的数据与真实数据的差异较大，模型效果相对较差。这种评价方法可有效评估生成文本的质量，并比较不同模型的性能。

4 实验结果

4.1 实验数据

CNN Dailymail 数据集来源于 CNN 和 Daily Mail 两新闻网站，涵盖了丰富的新闻报道内容，如政治、经济、科技、娱乐等领域[15]。数据集具有较高的质量和多样性，为研究者提供了一个理想的平台以研究和测试自然语言处理技术。

CNN Dailymail 数据集的文本已经经过预处理，其中每篇文本都包含标题、正文和摘要，其中摘要通常由几个关键句子组成，用以概括文本的主要内容。这种结构有利于研究者在文本摘要任务中进行有针对性的训练和测试，本研究中仅进行文本续写任务，故只考虑正文。

4.1 词性标注

4.1.1 NLTK

NLTK（Natural Language Toolkit）是一个用于自然语言处理的 Python 库，包含了各种文本处理和分析工具。其中，NLTK 的词性标注模块提供了对文本中单词进行词性标注的功能[16]。

在 NLTK 的词性标注模块中，使用的是基于隐马尔可夫模型[17]（Hidden Markov Model, HMM）的词性标注方法。该方法首先对标注语料库进行统计分析，

从中提取不同词性的出现频率和概率信息。然后，该方法将文本中的每个单词与不同的词性进行匹配，计算每种词性出现的概率，并选取概率最大的词性作为单词的标注[17]。在计算概率时，该方法会考虑前一个单词的标注信息，以提高标注的准确性和连续性。

表 1 词性表

词性	CC	NNS	…	NNPS
词性含义	连词	名词复数	…	专有名词复数

本研究的 POS(Part-Of-Speech 词性)基于 NLTK 工具中的词性标注序列，并人工将其余所有标点符号纳入“SEP”（separate）标签，以便于模型处理，可得到形同表 1 所示的 39 类词性序列。

分词处理结果

基于上述词性标注方法，本研究对训练集和测试集数据进行预处理，得到原始文本和与之对应的词性标注序列，形如表 2 所示的词性序列。

表 2 原始文本与对应词性

原始文本	对应词性
It ’ s official : U.S. President ...	PRP VBZ JJ SEP NNP NNP ...
(CNN) -- Usain Bolt rounded off ...	SEP NNP SEP SEP VBP NNP VBD RP ...

4.2 数据处理

本文采用三组规模不同的数据进行实验，旨在全面评估本模型在不同数据规模下的实际效果，并与原模型进行比较,数据规模如表 3 所示。实验基于三组规模不同的数据进行训练和评估，并采取 8:2 的比例对数据集进行随机分割，分别运行 10 次，以假正例率 *FPR* 为标准，去除最好和最差结果后取最高值、最低值和平均值作为结果数据。

表 3 数据规模

训练集大小	测试集大小	训练轮次
1600	400	5
16000	4000	1
32000	8000	1

4.3 实验结果

4.3.1 续写结果

通过表 4 中的两个例子，可以比较和分析 GPT-WP 模型和 GPT 模型产出的续写结果，总结 GPT-WP 模型的表现。

在第一个例子中,GPT-WP 模型生成的文本在语义上比原模型更加连贯。例如，“he is the first to have the best to play in his own world of his own world history”这一句，尽管有一些词语重复，但整体语义比 GPT 模型生成的“he is the first person in his career and he is the only one in two days to have the same”要更加清晰。第二个例子中的表现也是类似。

在第二个例子中，GPT-WP 模型生成的文本更好地理解和保留了输入文本的上下文。例如，续写文本中的“'dched the first day in a year and was a good person , and he has a great time in his own world”，反映了 John Schneider 对于 Jimmie 的评价，与输入文本的情感色彩相一致。而 GPT 模型生成的文本在这方面表现得相对较弱。

表 4 输入输出对比

输入文本	GPT 模型续写文本	GPT-WP 模型续写文本
The newspaper says is wearing latex gloves and carrying a black sack . He	has a great experience in his own world , but it is not the most of a big part in his career , and that he is the first person in his career and he is the only one in two days to have the same .	has a great experience in his own world , but it is not the most of a big part in his career , and that he is the first to have the best to play in his own world of his own world history , but it is not the most difficult time to get the best .
`` I laughed and learned more from Jimmie in one hour than from anyone else in a whole year , " co-star John Schneider , who played Bo Duke , said on Twitter .`` Give Uncle	' last week ' last month 'dched the first day in a year and was a good person , and he has a great time in his own world , but it is not the best part in his career , and that he is now in his own position , but it is a good time .	' last week ' last month 'dched the first day in a year and was a good person , and he has a great time in his own world , but it is not the best part in his career , and that he is now in his own position , but it is a very important way for us . "

总的来说，GPT-WP 模型在语义连贯性和上下文感知能力方面都表现出优越性，使得生成的文本更具有阅读性和理解性。然而，仍需注意模型在生成过程中存在重复词语和词组的问题，在一定程度上影响了输出文本的质量。在后续的模型训练和优化过程中，这是后续可以改进的问题。

4.4.2 评估结果

在三组不同规模的数据集上进行实验，结果如表 5 所示，最优结果已加粗。

表 5 数据规模

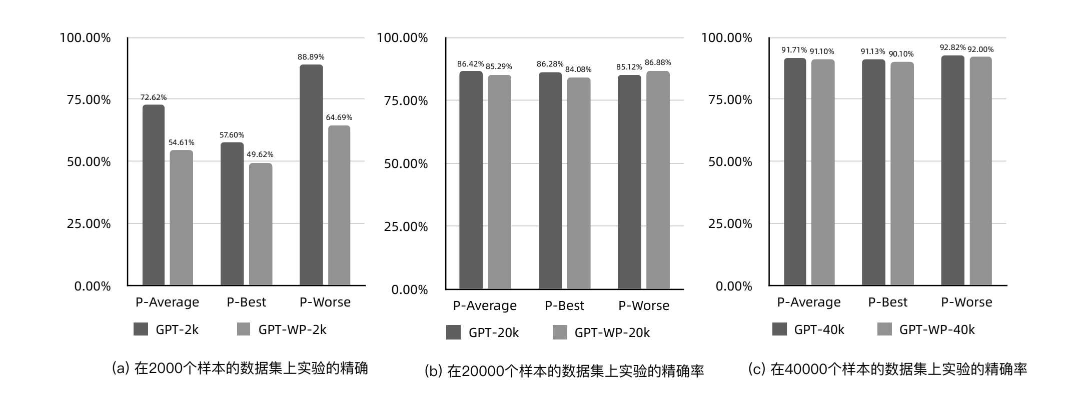
数据规模	模型	Acc			P			FPR		
		Average	Best	Worse	Average	Best	Worse	Average	Best	Worse
2k	GPT	72.91%	63.75%	**52.50%**	72.62%	57.60%	88.89%	26.42%	71.43%	0.49%
	GPT-WP	**59.06%**	**50.00%**	73.00%	**54.61%**	**49.62%**	**64.69%**	**80.60%**	**98.52%**	**52.71%**
20k	GPT	85.90%	88.18%	**82.13%**	86.42%	86.28%	**85.12%**	14.29%	15.37%	14.60%
	GPT-WP	**86.61%**	**86.50%**	87.43%	**85.29%**	**84.08%**	86.88%	**16.29%**	**18.20%**	**14.19%**
40k	GPT	92.92%	93.19%	93.27%	91.71%	91.13%	92.82%	8.52%	9.29%	7.24%
	GPT-WP	**92.88%**	**92.96%**	**92.04%**	**91.10%**	**90.10%**	**92.00%**	**9.27%**	**10.59%**	**7.99%**

1. 精确率

*P*度量的是正例（被识别为人工数据的文本）的样本中被预测为正例的比例，即 BERT 模型对人工数据的判别能力。

在 2000 个样本的数据集中，GPT 模型的平均精确率为 78.04%，而 GPT-WP 模型的精确率为 54.61%。这意味着在这个相对较小的数据集中，BERT 模型在区分由 GPT 和人工生成的数据时表现得更好，对分类模型的误导性较小，这表明 GPT 模型在文本生成方面的能力相对较弱。从最优和最差的精确率来看，GPT 模型分别达到了 57.60%和 88.89%的精确率，都超过了 GPT-WP 模型在最优和最差情况下的 49.62%和 64.69%的精确率。

然而，当数据集扩大到 20000 个样本时，两者的平均精确率都有所提升。GPT 模型的平均精确率提升至 86.42%，而 GPT-WP 模型的精确率提升至 85.29%。尽管两者的精确率非常接近，但 GPT 模型的精确率仍然略高。在最优和最差的精确率方面，两个模型的表现几乎相同。



(a) 在2000个样本的数据集上实验的精确 (b) 在20000个样本的数据集上实验的精确率 (c) 在40000个样本的数据集上实验的精确率

图 9 在不同数据集中的 BERT-tiny 模型精确率

当数据集进一步扩大到 40000 个样本时，GPT 模型的平均精确率提升至 91.71%，而 GPT-WP 模型的精确率为 91.10%，两者之间的差距缩小。在最优和最差的精确率方面，GPT 模型仍然稍高，其最优和最差的精确率分别为 91.13%和 92.82%，相比之下，GPT-WP 模型的最优和最差的精确率为 90.10%和 92.00%。

结合上述数据集的情况，无论是在较小或较大的数据集上，GPT 在精确率上的表现都高于 GPT-WP。这表明在 GPT 与人工模型组成的数据集上，BERT 模型对人工数据分类准确率更高，表明了 GPT 模型在文本生成的能力相对较弱。当数据集大小增加时，GPT 和 GPT-WP 的精确率都有所提升，但 GPT 模型的精确率仍高于 GPT-WP 模型。

2. 假正例率

假正例率 FPR 是一项关键的性能指标，主要度量了模型生成的文本被错误地识别为人工生成的比例。

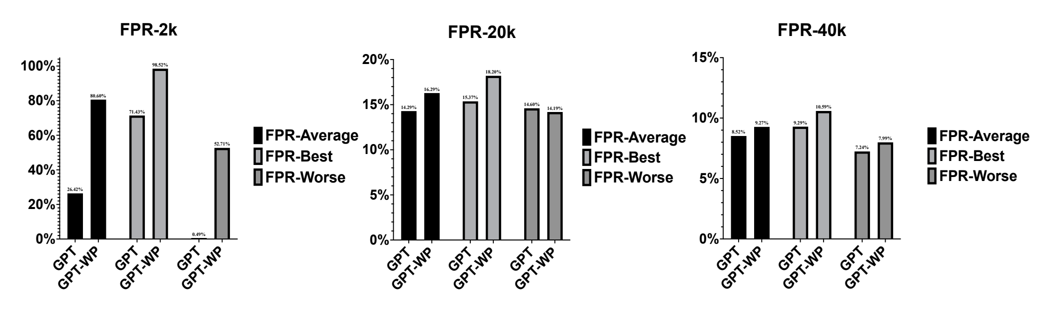


图 10 不同数据集上的 BERT-tiny 模型假正例率

本研究的主要推论在于，假正例率的增加意味着 BERT-tiny 模型将更多的模型生成数据判定为人工生成，这一现象表明模型生成的特征在更大程度上接近人工数据。这种接近程度使得模型生成的数据更难以被区分为机器生成，从而间接反映了模型生成文本的复杂性和逼真程度。因此，一个更高的假正例率表示生成的文本被错误判定为人工数据的比例增加，这可以被视为模型生成文本与人工文本接近程度的一个指标。在本研究中模型实验表现可以参考图 10。

在包含 2000 个样本数据集的实验中，GPT-WP 的平均假正例率为 80.60%，显著高于 GPT 模型的 26.42%。此外，无论是 98.52%的最佳假正例率还是 52.71%的最差假正例率上，GPT-WP 都明显超过了 GPT 模型。

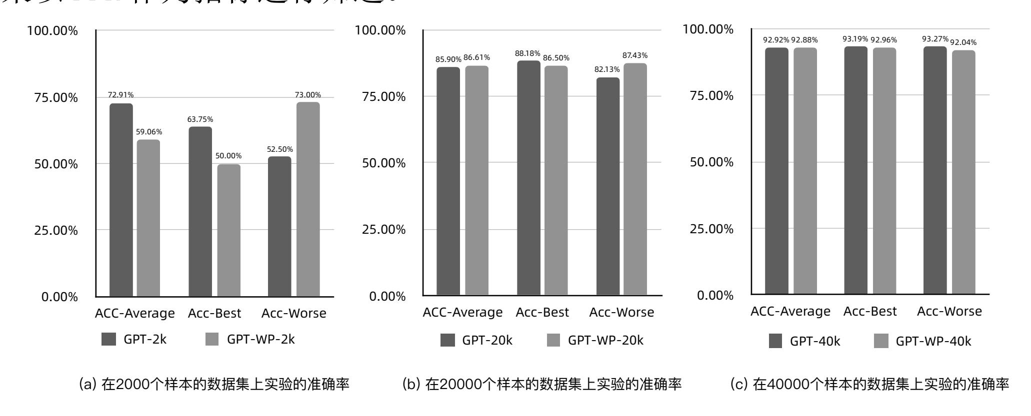
在包含 20000 个样本数据集的实验中，除最差假正例率略低以外，GPT-WP 模型在平均和最佳假正例率上显著高于 GPT 模型。在 40000 个样本的数据集中，GPT-WP 的平均假正例率为 9.27%，相比 GPT 模型的 8.52%，增长了 0.71 个百分

点。在最佳和最差假正例率上，GPT-WP 仍然优于 GPT 模型。这些数据进一步验证了 GPT-WP 模型在生成文本逼真性上优于 GPT 模型的假设。

更高的假正例率说明 GPT-WP 生成的文本更能混淆分类模型，使得 BERT 模型更难以识别出这些文本是由机器生成的。综合以上三个数据集的实验数据，本研究推断 GPT-WP 模型在生成文本的能力优于 GPT 模型。

3. 准确率

准确率Acc是衡量模型预测正确性的总体能力的关键指标。这是一个全局性指标，体现了 BERT 模型对所有类别预测的正确程度，在本实验场景中，类别即机器生成的文本和人工生成的文本，准确率表现如图 11 所示，其中最好和最差结果以 FPR 作为指标进行筛选。



(a) 在2000个样本的数据集上实验的准确率　(b) 在20000个样本的数据集上实验的准确率　(c) 在40000个样本的数据集上实验的准确率

图 11 不同数据集上的 BERT-tiny 模型准确率

在包含 2000 个样本数据集的实验中，GPT 模型的平均准确率为 72.91%显著高于 GPT-WP 模型的准确率平均值 59.06%。在最佳情况下，GPT 模型的准确率为 63.75%，而 GPT-WP 模型的最佳准确率为 50.00%，在最差情况下，GPT 模型为 52.50%，其情况相比 GPT-WP 较好。这一结果反映出 GPT-WP 模型能够对 BERT 模型的特征提取造成困难，使其准确率大幅下降。在包含 20000 个样本数据集的实验和包含 40000 个样本数据集的实验上，两个模型的准确率趋于相等，稳健性也更佳，但两者差距不大。

结合上述数据集的情况准确率的结果显示，在小型和大型数据集上，GPT 模型的预测准确性较高，而在中型数据集上差距则不大。在实验中，GPT-WP 模型相比 GPT 模型极值差异更大，这可能表明 BERT 模型在二分类任务上受到 GPT-WP 模型干扰较大，使其特征捕捉不稳定，准确率波动大，仍需结合精确率和假正例率进一步分析。

4. 实验总结

在针对包含 2000 个样本的数据集进行的分析中，当数据量相对较小时，BERT-tiny 模型经过 5 个 epoch 的训练后，其准确性与 GPT-WP 模型相比并无显著差异。然而，GPT-WP 模型在精确性的降低以及假正例率的提高方面表现更为显著。这一现象显示，与 GPT 模型生成的数据相比，GPT-WP 模型所生成的数据被错误地判断为人工数据的比例提高了，表明了 GPT-WP 模型所生成的数据质量相对原始 GPT 模型更高。

在对包含 20000 个样本的数据集和包含 40000 个样本的数据集的分析中，这两个数据集的规模较大，BERT-tiny 模型仅经过一个 epoch 的训练就能够获得较高的准确率和精确率，这表明 BERT-tiny 模型的内部参数得到了充分的学习，并且对于生成的文本特征和人工数据特征有着更好的理解。这一结果突出了参数量

规模效应的影响力。然而，即使在这种情况下，GPT-WP 模型生成的文本被误认为人工数据的比例仍然存在上升的趋势。因此，本研究在大样本数据集上进行的研究进一步证明，GPT-WP 模型相对于传统的 GPT 模型在文本生成方面具有优势。

综合来看，BERT-tiny 模型在处理由传统 GPT 模型生成的数据集时，其精确性和对负样本的识别能力都表现出优秀的效果，而 GPT-WP 模型生成的数据集在识别真正例样本方面表现出更好的结果，并且假正例率有较显著的提升。基于对三种规模实验的研究，本研究认为，与传统的 GPT 模型相比，GPT-WP 模型生成的文本在与人工数据相似性方面表现出一定的优势。

5 总结

本研究提出了一种基于词与词性的联合文本生成模型，GPT-WP,该模型由词级模型和词性级模型共同构成。通过将输入的文本及其词性序列分别编码并输入到两个模型中，使得生成的单词更加符合其语境语义。相较于传统的 GPT 模型，GPT-W 模型产生的文本更接近人工撰写的文本，在准确率、精确率及假正例率等评价指标方面表现出优势。

在评估方法方面，本研究提出并采用了一种创新性的方法，即利用 BERT 模型对生成文本的质量进行评估，并通过准确率、精确率和假正例率等指标对模型性能进行量化评价。与传统评估方法不同，本研究将原始数据与生成数据进行二分类任务，判断文本是否为人工编写，从而评估生成文本的质量。这种评估方法的创新之处在于，它能够在大规模生成任务中更加准确地评估生成文本的真实性和质量，以及更好地满足实际应用场景的需求。

综上，本研究提出的基于词汇与词性的联合文本生成模型在生成质量上取得一定的成果，结合创新的评估方法，为自然语言处理领域的文本生成任务提供一种新的解决方案，并为后续研究提供参考。

参考文献：**References**:

[1] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [J]. OpenAI blog, 2019, 1(8): 9.

[2] ZHANG X, LAPATA M. Chinese poetry generation with recurrent neural networks; proceedings of the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), F, 2014 [C].

[3] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [J]. Advances in neural information processing systems, 2013, 26.

[4] LOWE R, POW N, SERBAN I, et al. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems [J]. arXiv preprint arXiv:150608909, 2015.

[5] MELLISH C, DALE R. Evaluation in the context of natural language generation [J]. Computer Speech & Language, 1998, 12(4): 349-73.

[6] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735-80.

[7] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks [J]. Advances in neural information processing systems, 2014, 27.

[8] HOLTZMAN A, BUYS J, DU L, et al. The curious case of neural text degeneration [J]. arXiv preprint arXiv:190409751, 2019.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.
[10] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners [J]. Advances in neural information processing systems, 2020, 33: 1877-901.
[11] DEVLIN J, CHANG M-W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:181004805, 2018.
[12] MEDSKER L R, JAIN L. Recurrent neural networks [J]. Design and Applications, 2001, 5: 64-7.
[13] CHEN Y. Convolutional neural network for sentence classification [D]; University of Waterloo, 2015.
[14] ANDREJ, GROSS D, GOOD O. nanoGPT [Z]. GitHub repository. https://github.com/karpathy/nanoGPT; Github. 2023
[15] NALLAPATI R, ZHOU B, GULCEHRE C, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond [J]. arXiv preprint arXiv:160206023, 2016.
[16] LOPER E, BIRD S. Nltk: The natural language toolkit [J]. arXiv preprint cs/0205028, 2002.
[17] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition [J]. Proceedings of the IEEE, 1989, 77(2): 257-86.

(通讯作者: 李慧　E-mail:muxxsmu@outlook.com)

作者贡献声明*:

王蛟：提出研究思路，设计研究方案，实验并分析数据。
李慧：论文修订和研究过程指导

王蛟，首都师范大学教育学院本科生，邮箱：muxxsmu@outlook.com

李慧，首都师范大学教育学院副教授，硕士生导师，研究方向为人工智能与数据挖掘。联系电话：010-68980667。